Sound Multi-objective Feature Space Transformation for Clustering

Ingo Mierswa and Michael Wurst
Department of Computer Science
University of Dortmund

Abstract

In this work we propose a novel, generalized framework for feature space transformation in unsupervised knowledge discovery settings. Unsupervised feature space transformation inherently is a multi-objective optimization problem. In order to facilitate data exploration, transformations should increase the quality of the result and should still preserve as much of the original data set information as possible. We exemplify this relationship on the problem of data clustering. First, we show that existing approaches to multi-objective unsupervised feature selection do not pose the optimization problem in an appropriate way. Furthermore, using feature selection only is often not sufficient for real-world knowledge discovery tasks. We propose a new, generalized framework based on the idea of information preservation. This framework enables feature selection as well as feature construction for unsupervised learning. We compare our method against existing approaches on several real world data sets.

1 Introduction

Many knowledge discovery problems cannot be solved accurately by using the original feature space. This is due to several factors as noise, redundancy, sparsity or the fact that standard learning algorithms cannot represent necessary complex feature relationships. Both supervised and unsupervised knowledge discovery therefore depend on methods that transform the feature space in an appropriate way.

For supervised learning a set of labeled data points must be given. The learning method should merely find a function which *predicts* the label for unseen data points. The feature space transformation problem can be solved by finding a minimal feature space that maximizes the expected prediction accuracy.

Unsupervised machine learning differs essentially from supervised learning. The aim is usually rather to *describe* the data set and thus to automatically find inherent, natural patterns. Feature space transformation is important for unsupervised learning as well. Noise, sparsity and redundancy can hide the natural patterns in a data set, just as they can hide the relationship of the data points to a target function in supervised learning. There are several approaches that try to identify promising feature sets for unsupervised learning with respect to a task related criterion [Roth and Lange, 2003] but not with respect to a particular clustering algorithm. In addition, feature selection is a limited form of feature space transformation. It can for example not solve the problems of sparsity and feature interaction. Methods for feature space reduction can be applied to reduce noise and sparsity before applying unsupervised learning, e. g. Kernel-PCA [Schölkopf and Smola, 2002]. However, selecting appropriate parameters for new data sets is non trivial. This is especially problematic as such methods lead to feature spaces that are hard to interprete and resulting patterns are even harder to analyze. This is of course a clear conflict to the main target of cluster analysis.

The main limitation of these approaches is, however, that they do not reflect that feature space transformation for unsupervised learning inherently is a multi-objective optimization problem. Multi-objective problems are defined by several conflicting goals leading to the notion of Pareto optimal solutions. Several multi-objective wrapper approaches for unsupervised feature selection were proposed in [Kim *et al.*, 2000; 2002; Morita *et al.*, 2003]. These approaches minimize the number of features. Simultaneously, the quality of the identified patterns should be maximized. This idea is directly transferred from supervised multi-objective feature selection [Emmanouilidis *et al.*, 2000]. Figure 1 depicts the resulting Pareto front for a supervised feature selection problem. The used data set consists of 10 features necessary for classification and 10 additional noise features. It can clearly be seen that almost the complete range of solutions is covered, ranging from solutions containing only one feature and providing a small classification accuracy to a solution consisting of nine features with the highest accuracy. Adding the last non-noise feature or even noisy features would not lead to an improvement of accuracy and would therefore not lead to further Pareto optimal solutions. Hence, the Pareto front can not only be used as a feature selection method but also as a feature ranking method for the selected features.

While the basic idea of these approaches is very promising, they are limited in two points. First, minimizing the number of features just as in supervised learning is not robust for the unsupervised setting. Under very weak assumptions the set of Pareto optimal solutions collapses into one singular point that represents a trivial solution. Second, merely selecting features is not sufficient for many data mining tasks. In order to deal with problems as sparsity, new features must be constructed. We propose a new, generalized framework that approaches both problems. The quality of the resulting patterns should be optimized while the original feature space should be transformed as little as possible. Both goals are clearly conflicting as will be discussed in this paper.

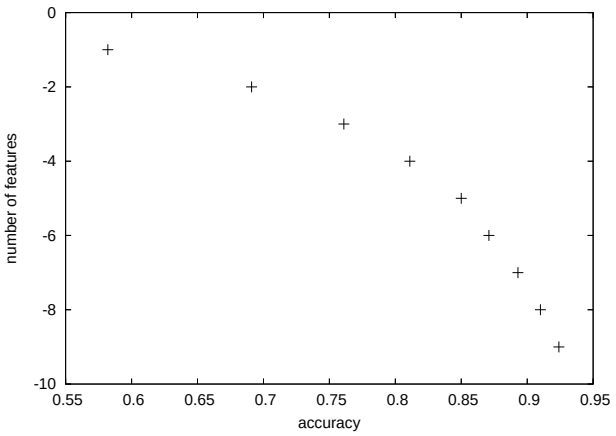


Figure 1: A typical Pareto front for supervised multi-objective feature selection.

1.1 Outline

In section 2 we will discuss existing approaches for multi-objective unsupervised feature selection for different cluster evaluation measures. Although the transfer from supervised learning is an appealing idea, we will show that these approaches do not lead to complete Pareto fronts for this type of problem. In section 3 we will discuss how simply changing the optimization direction for one of the criteria leads to a natural multi-objective optimization problem which will be solved by means of evolutionary algorithms. Finally, section 4 enriches the proposed framework by incorporating feature construction as well. Section 5 presents results on several artificial and real-world data sets and compares the discussed approaches. Section 6 concludes this paper.

2 Multi-objective feature selection for clustering

We will discuss in the next two sections why multi-objective optimization is a natural choice for selecting appropriate feature subsets for clustering problems. A straightforward approach for this type of optimization problem is to simultaneously optimize conflicting criteria by transforming the problem into a single-objective optimization problem. This leads to a set of user parameters which have to be defined in order to weight the criteria. However, in the clustering setting the user has no idea of criteria weights and, furthermore, there exist no simple decision about correct or wrong clusterings. Such a decision would totally depend on the amount of information the user can obtain from different clusterings. We try to maintain as much information as possible and aim at finding all solutions which are optimal for arbitrary criteria weight vectors. These solutions are called *Pareto-optimal*. The multi-objective search space of a maximization problem is subject to a partial order:

Definition 1 *A solution* a dominates *a solution* b *(written as* $a \succ b$*) if for the* p *criteria* r_i *the following is true:*

$$\forall i \in \{1,\ldots,p\} : r_i(a) \geq r_i(b) \quad \wedge \qquad (1)$$
$$\exists i \in \{1,\ldots,p\} : r_i(a) > r_i(b)$$

Our selection scheme needs to decide if a solution is dominated by a set B of solutions. We define:

Definition 2 *A solution* a *is* non-dominated *by a set of solutions* B *if* $\nexists b \in B : b \succ a$.

Now we are able to define what we mean with Pareto-optimal solutions:

Definition 3 *A solution* a *is* Pareto-optimal *if* a *is non-dominated by the complete solution space.*

The usual approach for multi-objective problems are evolutionary algorithms which can optimize more than one target function by introducing special selection operators [Coello Coello, 1999]. Traditional approaches in the field of mathematical programming must be applied more than once for multi-objective optimization [Yu and Zeleny, 1975]. Due to the population based approach of evolutionary algorithms a broad selection of Pareto-optimal solutions can be found during one run. The user can select one of these solutions after optimization. Additionally, multi-objective evolutionary algorithms do not strongly depend on form and continuity of the Pareto-optimal set [Coello Coello, 1999]. We will see in Section 5 that for clustering with non-normalized optimization criteria the Pareto front is neither nicely shaped nor continuous.

A basic condition to pose an multi-objective optimization problem properly is that the described criteria are actually in conflict to each other. By improving on one criterion, we cannot simultaneously improve on the other criteria. Only problems for which this condition holds are sound and can be solved by multi-objective optimization.

The current state of the art for multi-objective unsupervised feature selection is represented by the work initially described in [Kim *et al.*, 2000; 2002] and [Morita *et al.*, 2003]. In the following, we will describe both approaches and show that they are both limited in several ways. These limitations are a result of the way the multi-objective optimization problem is posed.

The corresponding methods all employ a wrapper approach. They subsequently apply a clustering scheme, e.g. k-means, to different feature subsets and evaluate the result with respect to several criteria. In [Kim *et al.*, 2002] four performance criteria for k-means clustering are used[1]. The first one is a variant of within cluster distance W that is normalized by the number of features

$$W_{norm} = \frac{1}{M} W \text{ with } W = \sum_{k=1}^{K} \sum_{x_i \in C_k} \sum_{m=1}^{M} (x_{im} - c_{km})^2 \qquad (2)$$

where c_{km} as the m-th value of the centroid of cluster C_k and x_{im} is the m-th value of the example x_i. The centroid is the point with the smallest distance to all points in C_k. A variant of between cluster distance is used as a second measure. However, this measure behaves essentially in the same way as W_{norm} (minimizing within cluster distance is equivalent to maximizing between cluster distance [Hastie *et al.*, 2001]). The third measure represents the number of clusters K which should be minimized. The last measure captures the number of features nf that should be minimized as well.

In the following theorem we show that for a given number of clusters K minimizing W_{norm} and the number of features leads to exactly one Pareto optimal point. This optimal point always selects one single feature from the dataset, in particular the one that leads to a minimal loss with respect to the used clustering performance criterion.

Theorem 1 *Minimizing* W_{norm} *and the number of features* nf *leads to one single Pareto optimal point.*

[1] In the original work all criteria are normalized by a constant. This, however, has no influence on Pareto optimality.

Proof: For W_{norm} we can denote the loss of an individual feature m as

$$a_m = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_{im} - c_{km})^2 \quad (3)$$

In order to minimize the number of features selecting only one feature is optimal. We show that always

$$W_{norm} \geq \min_{1 \leq m \leq M} \{a_m\}. \quad (4)$$

That means that the performance can only decrease by adding any feature but the one that optimizes a_m. It can easily be seen that

$$W_{norm} = \frac{1}{M} \sum_{m=1}^{M} a_m \quad (5)$$

$$\geq \frac{1}{M} \sum_{m=1}^{M} \min_{1 \leq m \leq M} \{a_m\} \quad (6)$$

$$= \min_{1 \leq m \leq M} \{a_m\} \quad (7)$$

Hence, using W_{norm} for optimization is not a well suited approach for feature selection in clustering problems as it leads to trivial solutions. Simultaneously minimizing W_{norm} and the number of features nf leads to one single Pareto optimal point. Therefore, selecting a single feature only is always the best solution for both criteria. The Pareto set collapses into a single solution. A similar proof can be given for normalized between cluster distance.

In [Morita *et al.*, 2003] a normalized variant of the DBIndex [Davies and Bouldin, 1979] is proposed as alternative performance criterion to W_{norm}, hence

$$DB_{norm} = \frac{1}{M} DB \quad (8)$$

$$\text{with } DB = \frac{1}{K} \sum_{k=1}^{K} \max_{k,l \neq k} \left\{ \frac{S_k + S_l}{d(c_{km}, c_{lm})} \right\}$$

where d is the Euclidean distance and S_k and S_l are the average within cluster distances for cluster C_k and C_l respectively which is defined as

$$S_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} d(x_i, c_k). \quad (9)$$

This approach is better suited, as the DBIndex is normalized with respect to the feature space. However, this criterion is still very sensitive. If the feature set contains for example a real valued feature that takes discrete values only, then choosing this one feature is again Pareto optimal, similar to the case of W_{norm}. However, this one feature does almost certainly not represent the complete dataset in the descriptive sense mentioned in the introduction. In section 5, we will see several examples for which the Pareto set collapses into a single trivial solution even for normalized DBIndex or, at least, for which the resulting Pareto sets do not cover the complete range of possible feature subsets. The same applies for other recently proposed normalization schemes [Handl and Knowles, 2006] which basically just increase the weighting factor between the number of features and the cluster evaluation measure.

The major problem of these approaches is that they do not pose the problem correctly from the point of view of multi-objective optimization. In the next section we give an alternative problem formulation that solves the described difficulties.

3 Information preserving feature selection

In the last section, we discussed several quality measurements for clustering schemes. In the following, we assume that all criteria should be maximized during feature selection. In contrast to the existing approaches discussed in section 2 we do not *minimize* the number nf of features but *maximize* it. This change of the optimization direction directly follows from Theorem 1. Although maximizing the number of features during feature selection might sound surprising at first, this paradigm change can be motivated by the aim of unsupervised learning: the search for descriptive patterns. Maximizing the number of features prevents the algorithm from selecting trivial solutions and leads to more complete Pareto sets of diversive natural clusterings. The fitness is evaluated by performing a clustering scheme on the reduced feature sets. We use DB as quality criterion. Since there is a natural competition between maximizing the number of features nf and the cluster criterion we do not need to apply an artificial normalization factor as in DB_{norm}.

We use NSGA-II as a multi-objective feature selection wrapper [Deb *et al.*, 2002]. NSGA-II employs a selection technique which first sorts all individuals into levels of non-domination. Individuals from the first levels are added to the next generation until the desired population size is reached. Before individuals are added from the last possible level, this level is sorted with respect to the crowding distance in order to preserve diversity in the population.

Individuals are bit vectors of length M indicating if a feature should be selected or not. The population size is set to $2M$, the maximal number of generations is 1000. A bit flip mutation is performed with probability $1/M$ and uniform crossover with probability 0.9.

4 Information preserving feature aggregation

Merely selecting features is often not sufficient. First, sparse data is a severe problem in many applications areas. For text clustering, generalizing terms by adding superordinate terms can significantly improve the quality of the result [Hotho *et al.*, 2003]. The same holds for association rule mining. Adding generalized features, which combine individual items to classes, enables the algorithm to find patterns, which would not be valid in the original data space [Srikant and Agrawal, 1995]. Second, many datasets contain features produced by similar underlying processes, e.g. time series data. Popular preprocessing approaches as moving average replace neighboring values by a generalized value exploiting the assumption that neighbors are similar.

In the following, we present a general formalism for feature aggregation. This formalism is a straightforward generalization of the feature selection framework presented in the last section and should fulfill several requirements. First, the constructed feature space should be easily interpretable in order to allow for a quick inspection of the results. Second, the optimization problem should be posed in a way that it can be solved efficiently. Third, as for selection, trivial solutions must be avoided. Finally, the aggregation value should deviate as little as possible from both given feature values. This last property again is necessary in order to properly define a multi-objective feature space transformation similar to the mere selection problem discussed in the last section.

Definition 4 *Let X denote the data set and X_r, X_s, and X_t single features. A* feature aggregation function *is a function $f : X_r \times X_s \rightarrow X_t$ that maps two features to a new feature.*

Please note that the newly aggregated feature replaces the arguments. In the following, we state formal conditions that an aggregation function should fulfill to meet the points mentioned above. As point of departure, we use the concept of *t-conorms*, which naturally captures the notion of disjunctive value combinations. T-conorms are a class of theoretically and empirically established generic aggregation functions. They are a natural extension of disjunctions for continuous values. Disjunctions have proven to be essential for many data mining applications, e.g. for generalized association rules [Srikant and Agrawal, 1995].

Definition 5 *A function is a* t-conorm *if it fulfills the following constraints:*

1. *Boundary condition:*
$$f(X, 0) = X \tag{10}$$
2. *Commutativity:*
$$f(X_r, X_s) = f(X_s, X_r) \tag{11}$$
3. *Associativity:*
$$f(f(X_r, X_s), X_t) = f(X_r, f(X_s, X_t)) \tag{12}$$
4. *Monotonicity:*
$$X_r \geq X_s \Rightarrow f(X_r, X_t) \geq f(X_s, X_t) \tag{13}$$

Associativity and commutativity ensure that the feature aggregation is order independent, thus that the order in which features are aggregated does not have an influence on the result. This is of course not only desirable for disjunctive aggregations but also considerably reduces the search space and leads to results that are easier to interprete, as the system produces sets of features instead of trees. The boundary condition ensures that the aggregation follows the notion of a disjunctive merging (in contrast to $f(X, 0) = 0$, which would describe a conjunctive aggregation). Monotonicity preserves the ordinal information in the data.

Although t-conorms already can be used for disjunctive aggregations, they are, however, not sufficient to capture the notion of a minimal deviation. For example, it would still be possible that $f(x, x) \neq x$. Thus, even if both features have the same value, the resulting value could be different. This clearly violates the concept of merging two features and altering them minimally as stated above. We therefore add an additional constraint that excludes such functions:

Condition 1 *A function is an* information preserving t-conorm *if it is a t-conorm and fulfills the following minimal deviation condition:*

$$\begin{aligned} &\forall x, y \in X : \neg\exists f'(x, y) : \\ &|f'(x, y) - x| + |f'(x, y) - y| < \\ &|f(x, y) - x| + |f(x, y) - y| \end{aligned} \tag{14}$$

This condition states that the aggregation function should always yield a merged value that has a minimal deviation from both original values. In the following, we show that from the t-conorm conditions and Condition 1, two important properties can analytically be derived. The first property was discussed before and states that the aggregation result for equal values should again be the value itself. Otherwise, users would not be able to understand the meaning of aggregated features and it would not be possible to guarantee that the aggregated features are in any way similar to the original features. This property is called *idempotence* and directly follows from Condition 1:

Lemma 1 *Each information preserving t-conorm fulfills* idempotence, *i.e. $f(x, x) = x$ (proof trivial).*

Still, there might be a problem for non-equal values if the aggregated value would differ too much from the original values. In order to prevent the aggregation function to generate arbitrary values we set a last condition for aggregation functions:

Condition 2 *A function fulfills* domain preservation *iff $min(x, y) \leq f(x, y) \leq max(x, y)$.*

Thus the merged value must be in the domain spanned by the input values. We can show that Condition 2 can directly be followed from Condition 1:

Theorem 2 *A t-conorm $f(x, y)$ that fulfills Condition 1 also fulfills domain preservation.*

Proof: For $x = y$ the condition is trivially violated. We have to prove four cases and assume that $f(x, y) > max(x, y)$ and $x > y$. Then:

$$\begin{aligned} &|f(x, y) - x| + |f(x, y) - y| = \\ &(f(x, y) - x) + (f(x, y) - y) > \\ &(f(x, y) - x) + (x - y) \geq (x - y) = \\ &|max(x, y) - x| + |max(x, y) - y| \end{aligned} \tag{15}$$

Thus condition 1 is violated. The other cases can be shown analogously.

Together with Lemma 1, this theorem states that Condition 2 is a sufficient condition for information preserving t-conorms (Condition 1). Moreover, the above conditions constrain the set of possible aggregation functions to exactly a single one, the maximum function:

Corollary 1 *The maximum function is the only aggregation function fulfilling the information preserving t-conorm condition.*

Proof: It can be shown that for all t-conorms $f(x, y)$ the following holds: $max(x, y) \leq f(x, y)$ (proof trivial). On the other hand, the domain preservation conditions requires that $f(x, y) \leq max(x, y)$, hence $f(x, y) = max(x, y)$.

Given the aggregation function, we still need to extend the performance measure proposed above, to capture feature aggregation as well. We have seen that for mere feature selection the number nf of selected features is sufficient for measuring the degree of feature space preservation. One of the surprising results of this work is that this number should be maximized instead of minimized in the unsupervised setting. We want to extend the proposed framework in a way that feature selection is a special case of the more generic feature space transformation setting. We give two conditions which must be fulfilled by this generalized cost measure:

Condition 3 *Let nf_o be the number of selected original, i.e. non-aggregated, features. Let nf_a be the number of aggregated features in the transformed feature set. For an unsupervised feature space transformation measure nf the following must hold:*

Abba.	properties	N	M	noise	σ_o	σ_n	K	Results
GRID	equidistant values	3125	5	0	–	–	0	(a) and (b)
RANDOM	uniformly distributed	500	10	10	–	∞	0	(c) and (d)
GM	Gaussian mixture	1000	15	10	0.5	0.5	16	(e) and (f)
GM-L	Gaussian mixture	100000	15	10	0.5	0.5	16	(g) and (h)
IRIS	Iris without noise	150	4	0	0.8	–	3	(i) and (j)
IRIS-NN	Iris with nominal noise	150	5	1	0.8	0.01	3	(k) and (l)
IRIS-GN	Iris with Gaussian noise	150	14	10	0.8	0.8	3	(m) and (n)
WPBC	WPBC without noise	198	34	0	33.2	–	?	(o) and (p)

Table 1: The used data sets for unsupervised feature selection. The first column summarizes the used abbreviations, the second describes the data set. N is the total number of examples, M the number of features. *Noise* defines how many features of M where explicitly added noise features. The next columns define the mean standard deviation of the original (σ_o) and the noise features (σ_n). The column K indicates the number of clusters if known. The last column indicates which Pareto sets were found with both approaches.

abbr.	properties	N	M	K	Results
IRIS-M	Iris data set with divided features	150	8	3	(a)
KDDCUP	quantum physics data (KDD cup 2004)	5000	78	2	(b)
NEWS	articles from three newsgroups	3000	1052	3	(c)

Table 2: The used data sets for unsupervised feature space transformation.

1. *$nf = nf_o$ if the feature set does not contain any aggregated features.*
2. *Every aggregation must lead to a loss of $-a$ with $a > 0$.*

In the following, we will assume $a = 1$. A very simple measure fulfilling these conditions is given by $nf = nf_o + nf_a$. If no features were aggregated nf_a is 0 and $nf = nf_o$. Since all aggregation functions must replace the input features, aggregating two original base features reduces nf_o by 2 and increases nf_a by 1. Hence nf is totally increased by 1. The same applies in the case of two already aggregated features or in the case of a merge of one base feature with an already aggregated features. Hence, every aggregation leads to the same loss of -1. Just as for the mere feature selection case the number nf should be maximized in order to ensure minimal deviation and thus a set of conflicting criteria. This again leads to a proper definition of a multi-objective optimization problem even for the unsupervised feature transfer setting.

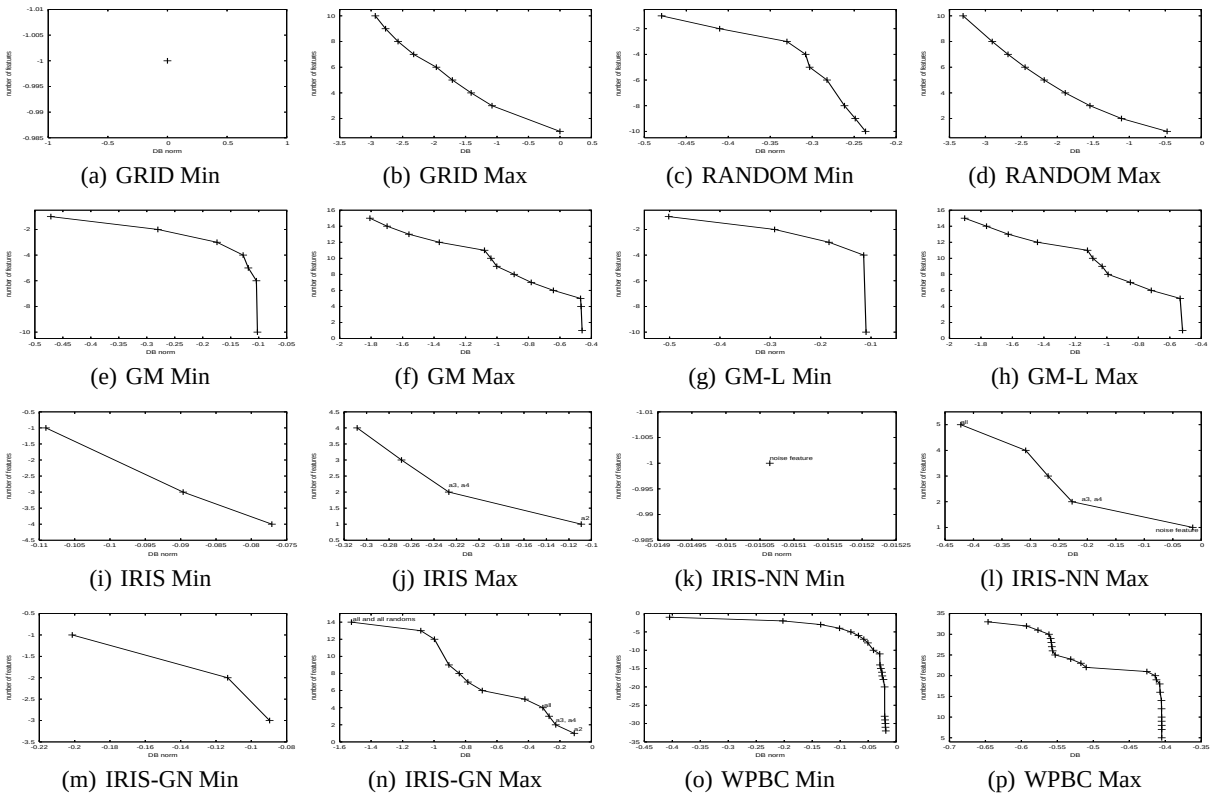
Allowing the aggregation of features induce a representation change for the individuals of the evolutionary algorithm. Individuals are still represented by vectors $\vec{v}$ of length M. In contrast to the feature selection case, each coefficient of this vector is a number $v_i \in [-1, \max(v_1, \ldots, v_M)]$. This number v_i represents the state of the i-th feature. -1 means that the feature is not selected at all. 0 means that the feature is used in its original form. Any number greater than 0 means that the feature should be aggregated with other features with the same number. This ensures that each feature is used at most once in the complete set. The mutation operator performs an uniformly distributed random change of each coefficient in the interval $[-1, max(v_1, \ldots, v_M) + 1]$. This mutation is performed with probability $1/M$ for each coefficient. The other algorithm parameters are the same as in the special case of feature selection.

One important property of our approach is that the number of features is strictly monotonically decreasing. This is important for the efficiency of the proposed method, as decreasing the number of features will decrease the runtime of the inner clustering algorithm. In contrast to other feature construction approaches the used vector representation also ensures that the amount of memory is restricted to the start individual size. Therefore, our approach can also be used for large scale unsupervised feature selection and aggregation and is feasible even for large data sets with many features.

5 Evaluation

We first compare our approach to existing approaches for multi-objective unsupervised feature selection. We then analyze the properties of our generalized feature space transformation on several synthetic and real-world data sets. The essential requirement is that the resulting Pareto sets are as broad as possible. The worst case is a Pareto front that collapses into a single point.

In order to measure the effect of the artificial normalization factor necessary for the existing feature set minimization approach, we applied the algorithms on a grid data set (GRID) and a random data set (RANDOM) containing only white noise. Another data set (GM) consisting of 16 Gaussian clusters with random standard deviations between 0.0 and 1.0 in five dimensions was created. This data set was enriched with ten additional single Gaussian noise features with average standard deviation 0.5. The same data set but with 100000 examples was created in order to check if our approach is feasible with respect to large data set sizes (GM-L). We also applied both algorithms on two clustering benchmark datasets, namely the IRIS data set [Fisher, 1936] and the WPBC (Wisconsin Prognostic Breast Cancer) data set [Wolberg *et al.*, 1995]. These data sets were also used by [Kim *et al.*, 2000; 2002; Morita *et al.*, 2003] for evaluation. The WPBC data set is especially interesting because of many redundant features.



(a) GRID Min (b) GRID Max (c) RANDOM Min (d) RANDOM Max

(e) GM Min (f) GM Max (g) GM-L Min (h) GM-L Max

(i) IRIS Min (j) IRIS Max (k) IRIS-NN Min (l) IRIS-NN Max

(m) IRIS-GN Min (n) IRIS-GN Max (o) WPBC Min (p) WPBC Max

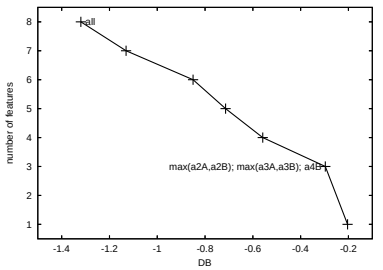
Figure 2: The Pareto fronts for all data sets. The left result for each dataset is achieved by the approach discussed in section 2 for a normalized DBIndex (nDB) against minimizing nf. It can clearly be seen that these results are not as complete and that kinks are covered by the artificial $1/x$ structure. The results on the right are achieved by our maximization approach, thus non-normalized DBIndex (DB) against maximizing nf.

This allows us to check how well both approaches are able to cope with redundancy. Table 1 summarizes the properties of all data sets.

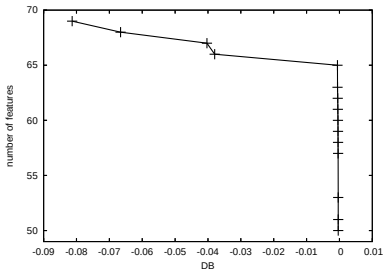
All experiments were performed with the free machine learning environment YALE [Mierswa *et al.*, 2006]. It should be noted that in most cases the population converges to the final front after less than 20 generations. The NSGA-II selection was able to sustain the found solution until the end of optimization. Figure 2 shows all Pareto sets for the simultaneous optimization of the used cluster criterion and the feature set size. The achieved performance (DB or DB_{norm}) is depicted on the x-axis, the number of features nf is depicted on the y-axis. In case of the minimization approach, the number of features is multiplied with -1 for optimization. In order to turn the problem into a full maximization problem, we also multiplied DBIndex with -1.

One might ask why the comparison plots have different scales and variables. Former experiments have shown that both approaches are able to deliver Pareto-optimal solutions independently of the used scale. However, a scale based comparison alone is not applicable in order to decide which Pareto sets are superior. Hence, other criteria like completeness or shape of the fronts must be taken into account. One of the insights of our work is that a normalization factor, as proposed by other authors, is not necessary if the number of (original) features is maximized. Since the normalization induces an additional artificial competition *and* covers inherent structures in the Pareto sets, we decided to plot the results for non-normalized DB. For normalized DB_{norm}, which would lead to the same scales for both approaches, our approach simply produces a super set of solutions. Moreover, if non-normalized DB is used for the formerly proposed minimization approach, the Pareto fronts collapse in almost all cases.

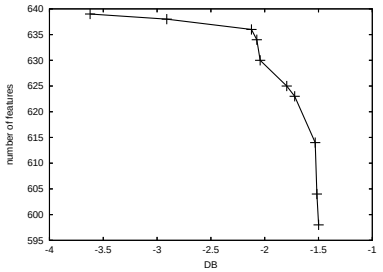
It can clearly be seen that in all cases the Pareto sets provided by our approach contain more points than the results of the normalized minimization approach. If there is only one feature with a relative small standard deviation, the Pareto set of the minimization approach will still collapse (GRID and IRIS-NN) even for normalized DBIndex. Of course, well-defined multi-objective solutions should be able to deliver the complete Pareto front including more than only this one trivial solution. Moreover, the normalization factor $1/x$ introduces a convex front although there is nothing to optimize at all. This effect can be seen for the RANDOM data set, where the minimization approach finds a convex Pareto front while the front provided by our approach is still linear. For the Gaussian mixture clusters (GM), again our approach is able to deliver a broader Pareto front including some kinks. These kinks can be used as hint for interesting regions of the Pareto front easing the final selection of solutions. The best clustering result for the minimization approach was the feature set at the right end of the Pareto front containing 10 features. The found clusterings for this feature set, however, did not correspond to the original clusterings at all. On the other hand, our approach was able to find the correct feature set consisting of 5 features providing 12 of the 16 original clusters (the first kink seen from the right end). It can also be seen that the main structure of the Pareto front remains with respect to the sample size. Applying our approach on GM-L with 100000 examples was feasible and delivers similar results.



(a) IRIS-M



(b) KDDCUP



(c) NEWS

Figure 3: The Pareto fronts delivered by the unsupervised multi-objective feature aggregation experiments. The Pareto sets still cover the complete range of possible solutions, e.g. from 1 until 8 features for the IRIS-M data set. Additionally, features were only aggregated if this combination was necessary.

For both the normal IRIS data set and the IRIS-GN data set enriched with noise features the proposed approach finds the complete Pareto set including the correct clusterings while the minimization approach was only able to find a small number of feature subsets. Correct clustering are depicted by small labels indicating the used feature set. For the IRIS data set without noise, the minimization approach was also able to deliver a feature subset of 3 features providing the correct clustering. In contrast to our approach, the minimization approach collapses (IRIS-NN) or was not able to deliver the correct clustering (IRIS-GN). Since it is not clear which clustering is "correct" beforehand, the user should be able to select from the complete Pareto front delivered by our approach. The same conclusion applies for the other real-world data set WPBC.

In addition to the comparison between our approach and the formerly proposed approach, we also applied the new unsupervised feature aggregation algorithm on one semi-synthetic and two real-world data sets. The properties of these data sets are summarized in Table 2. For the data set IRIS-M we divided the values of the four Iris features into two parts A and B resulting in a total of eight features. For each of the original feature values we randomly select one of the new features as target, the other feature value

Data set	$\|P\|$		$\|C\|$		F	
	min	max	min	max	min	max
GRID	1	**9**	0	0	?	?
RANDOM	9	9	0	0	–	–
GM	7	**13**	0	**12**	no	**yes**
GM-L	5	**12**	0	**11**	no	**yes**
IRIS	3	**4**	3	3	yes	yes
IRIS-NN	1	**5**	0	**3**	no	**yes**
IRIS-GN	3	**12**	0	**3**	no	**yes**
WPBC	21	**24**	?	?	?	?
IRIS-M	–	**7**	–	**3**	–	**yes**
KDDCUP	–	**15**	–	**2**	–	–
NEWS	–	**10**	–	**3**	–	–

Table 3: Comparison of the results for the minimization approach and the proposed non-normalized maximization approach. Better values are indicated with a bold font.

is set to a random value between 0 and the current value. This way the complete original information can only be reconstructed by aggregating the correct features. The KDDCUP data set consists of a stratified sample of 5000 examples drawn from the quantum physics data of the KDD cup 2004. For the data set NEWS, we combined three newsgroups of the well known 20-newsgroups data set which results in 3000 examples.

Figure 3 shows the results for unsupervised feature space transformation. For the IRIS-M data set, the complete range of solutions is covered by the resulting Pareto set and the necessary features $a2$ and $a3$ were reconstructed by aggregation. At this point, the known clustering of the IRIS data set was found by our approach (again depicted by a label indicating the used feature set). For KDDCUP, a clear kink can be seen indicating redundant features which are aggregated in the lower part of the vertical line. For both the KDDCUP data set and the NEWS data set two respectively three clusters were found covering large parts of the original classes. For all real-world data sets a broad range of the feature space is covered by the result which again supports our claim of robust and useful solutions.

Table 3 summarizes all results for both the mere feature selection case and the feature aggregation case. $|P|$ denotes the number of found Pareto points, $|C|$ indicates the number of found correct clusterings, and column F indicates if the correct feature set was found (if known). Better values are marked with a bold font. The hyphen indicates, that the approach can not be applied to this data set, the question mark indicates that the correct values are not known. It can clearly be seen that the new approach outperforms existing approaches in terms of Pareto front completeness, robustness, and ability to find correct clusterings.

6 Conclusion

We presented a novel multi-objective framework for feature space transformation in clustering settings which plays an important role in a wide variety of applications ranging from pattern recognition to customer relationship management and web search. Clustering is an inherently multi-objective problem. There is usually not one correct result as for supervised learning. Users rather explore the space

of results interactively to gain insight into the natural patterns within the data set.

Our work is based on previous work on multi-objective feature selection for clustering. We found, however, that existing approaches were limited in two points. First, they do not pose the optimization problem in a sound and robust way. We showed both analytically and empirically, that the corresponding sets of Pareto optimal solutions collapse to a single, trivial solution. We therefore proposed an approach that is based on the idea of information preservation. As much of the original data space should be preserved as possible, while the validity of the resulting clusters is optimized. We show that this approach yields complete and useful Pareto sets. The original feature set and clustering were found in all cases. These Pareto sets moreover show a strong inner structure which can be used to explore the set of solutions even more efficiently by inspecting only these interesting points.

In addition, merely selecting features is not sufficient in many settings. Especially the problems of sparse data and feature interactions cannot be properly solved by feature selection only. We extended our approach to allow for a limited form of feature construction as well. Aggregation is used to derive new features, that generalize over two or more features in the original data set. T-conorms are a class of theoretically and empirically established generic aggregation functions. They are a natural extension of disjunctions for continuous values, which have proven to be essential for many data mining applications, e.g. generalized association rules. A set of basic conditions limits feature aggregation to the t-conorm maximum which summarizes two features with minimal alteration. We show that even for feature aggregation our approach leads to robust Pareto sets. Our experiments supports this claim.

Also, our approach is very generic. As it essentially adopts a wrapper approach, it can be combined with a large variety of problems and algorithms. It is therefore easy to adapt to new problems and application domains. We successfully applied the proposed approach also for graph-based clusterings or density based clusterings or for other performance criteria [Handl and Knowles, 2006] without need for additional normalization.

References

[Coello Coello, 1999] C. A. Coello Coello. A comprehensive survey of evolutionary-based multiobjective optimization techniques. *Knowledge and Information Systems*, 1(3):129–156, 1999.

[Davies and Bouldin, 1979] D.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

[Deb *et al.*, 2002] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multi-objective genetic algorithm: Nsga-ii. Technical report, Kanpur Genetic Algorithms Laboratory, Indian Institute of Technology, 2002.

[Emmanouilidis *et al.*, 2000] C. Emmanouilidis, A. Hunter, and J. MacIntyre. A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. In *Proc. of the Congress on Evolutionary Computation (CEC)*, pages 309–316, 2000.

[Fisher, 1936] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[Handl and Knowles, 2006] J. Handl and J. Knowles. Feature subset selection in unsupervised learning via multiobjective optimization. *International Journal on Computational Intelligence Research*, 2006.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2001.

[Hotho *et al.*, 2003] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Proc. of the IEEE International Conference on Data Mining (ICDM 2003)*, 2003.

[Kim *et al.*, 2000] Y. Kim, W.N. Street, and F. Menczer. Feature selection in unsupervised learning via evolutionary search. In *Proc. of the 6th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 365–369, New York, NY, USA, 2000. ACM Press.

[Kim *et al.*, 2002] Y. Kim, W. N. Street, and F. Menczer. Evolutionary model selection in unsupervised learning. *Intelligent Data Analysis*, 6:531–556, 2002.

[Mierswa *et al.*, 2006] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. YALE: Rapid prototyping for complex data mining tasks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, 2006.

[Morita *et al.*, 2003] M. Morita, R. Sabourin, F. Bortolozzi, and C.Y. Suen. Unsupervised feature selection using multi-objective genetic algorithms for handwritten word recognition. In *Proc. of the 7th International Conference on Document Analysis and Recognition (ICDAR)*, 2003.

[Roth and Lange, 2003] V. Roth and T. Lange. Feature selection in clustering problems. In *Proc. of Neural Information Processing Systems (NIPS)*, 2003.

[Schölkopf and Smola, 2002] B. Schölkopf and A. J. Smola. *Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[Srikant and Agrawal, 1995] Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland*, pages 407–419. Morgan Kaufmann, 1995.

[Wolberg *et al.*, 1995] W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computer-derived nuclear "grade" and breast cancer prognosis. *Analytical and Quantitative Cytology and Histology*, 17:257–264, 1995.

[Yu and Zeleny, 1975] P. L. Yu and M. Zeleny. The set of all nondominated solutions in linear cases and a multicriteria Simplex method. *Journal of Mathematical Analysis and Applications*, 49:430–468, 1975.